International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 20s, 2025
https://theaspd.com/index.php

Review Of Deep Learning Models For Automatic Image Caption Generation

Mehzabeen Kaur[1], Harpreet Kaur[2]
[1]Ph.D Research Scholar, Department of Computer Science & Engineering, Punjabi University, Patiala, Punjab, India, mehzabeenkaur93@gmail.com
[2]Assistant Professor, Department of Computer Science & Engineering, Punjabi University, Patiala, Punjab, India, harpreet.cse@pbiu.ac.in

Abstract: *In the modern-day scenario, social media platform plays a major role in terms of image sharing and posting. A considerable number of images are uploaded on social media website like Facebook, WhatsApp, and Messenger in each day. Image captioning is the technique to describe an image into human understandable language. Deep neural networks and availability of efficient image captioning datasets helps to generate a suitable caption of the image in a faster way. In this review paper, latest advancement in image captioning techniques, deep neural networks, image captioning datasets and evaluation methods are presented. This paper not only meant to be review of image captioning rather it also discussed its strengths and limitations of most commonly approaches, datasets and evaluation metrics.*
Keyword*: Deep Learning, Image Captioning. Automated Image Captioning, Long Short-Term Memory (LSTM), Machine Learning.*

1. INTRODUCTION

While an image has "a thousand words", it is hardly ever practical to describe an image with so many. Instead, what is important in many areas is an adequate caption of an image which provides the main description. Image description is considered to be one of the intellectual challenging tasks now a days. A computer graphic method is proposed to recognise necessary objects and regions in photos on social media platform. Image captioning and description generation are the latest research topic comes with challenging for deep learning. A system is developed that can not only accurately label image regions but also scale to full image description for different applications. A hierarchical trained deep learning network is used to increase the fluidity and descriptive nature of the generated image captions [4],[9],[26]. The deep learning network operation performs in two different stages. Faster R-CNN achieves the initial proposal generation and regional description is translate into full image description by Recurrent Neural Network (RNN) based encoder-decoder structure. The proposed deep learning method can label scenes, object attributes, humans, and objects simultaneously. The regional proposal generation task is too noisy in deep learning image captioning and descriptive generation method. A comprehensive review of existing deep-learning-based image captioning techniques is presented in [44],[45],[48]. Various deep learning image captioning methods are analysed based on their operations like feature generation and multimodal stage.

2. Image Captioning

Image captioning is a process to generate the human like caption or description of the image in an artificial way. In this we provide the input as an image to the system and it classify the image and produce valid caption. In the first phase, image is recognised from different perspective and in second phase language model is used to present the meaningful caption of image (figure 1).

Figure 1: Examples of Image Captioning Outputs Generated by Deep Learning Models

2.1Types of Image Captioning

Several techniques have been proposed in the literature to help those people who had a low vision to understand the image data. Firstly, there existed template[13], [17], [28], [31], [37] and retrieval based image description approaches used for automatic image description generation. The proposed CNN-LSTM model provided strong platform for relational description in high-resolution images. The appropriate size of the attribute is the only limitation of CNN-LSTM model. The overfitting problem is minimized by using the deep CNN model [49]. Image captioning can be broadly divided into three categories i.e., retrieval-based captioning, template based captioning and deep learning-based captioning as shown in figure 2.

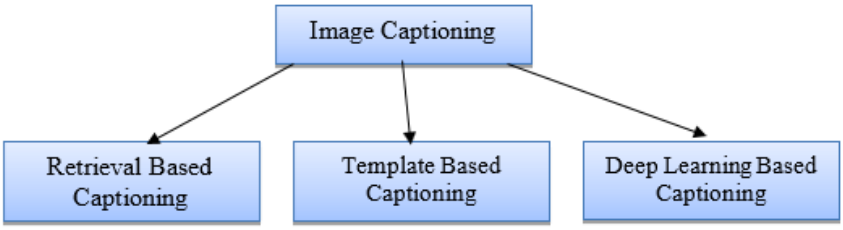

Figure 2: Types of different types of Image Captioning Approaches

- **Retrieval based methods:** it represents a description for a picture by accessing one or a group of word from a pre-defined sentence pool. The final description could either be a full meaningful English sentence which is predefined in an image dataset [1].
- **Template based methods**: it produces picture descriptions which is correct in the sense of linguistic and meaning. Basically, for using template-based description pre-defined fixed set of graphic ideas required to be perceived first. After that it is connected with linguistic rules [1].
- **Deep learning-based captioning** allows computer systems to recognize images for mainly education purposes, sentiment analysis, an aid for the visibly impaired, etc. The model must be accurate enough to understand the various relations between various objects, and express that in a correct semantic manner in natural language [1].

2.2 Image Captioning and Machine Learning

Machine learning (ML) or intelligent retrieval is a part of artificial intelligence (AI). The main target of ML is basically is to recognize the formation of facts and adjust that information into developed models (figure 3) which can be readable and used by individuals. However, ML is an area of computer science with is different form conventional computing approach. ML approach permit for system to learn on facts inputs and use arithmetical examination, if the results of figures come within a target. Supervised and unsupervised based learning approaches are used in machine learning for classification of the images. It uses the following techniques:

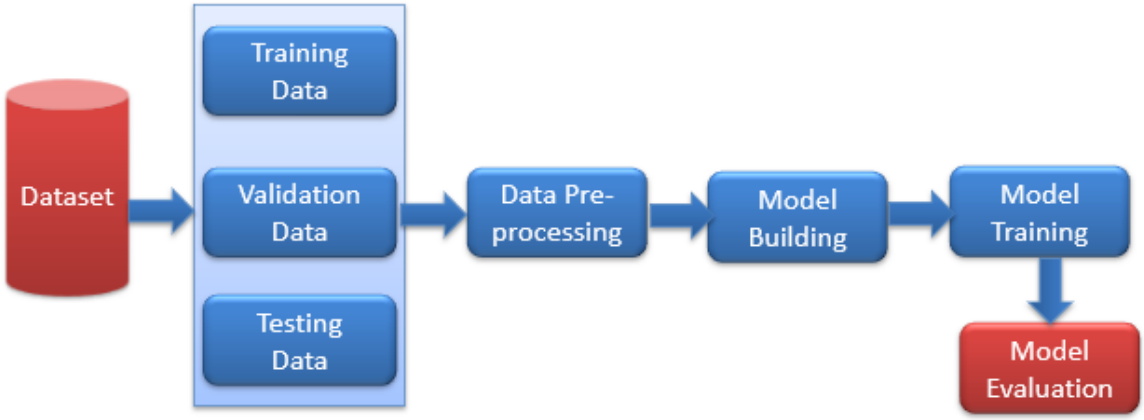

Figure 3: Architecture of Image Captioning Model Integrating CNN for Visual Feature Extraction and RNN/LSTM for Generating Descriptive Textual Captions

- **Regression:** Regression based approach are widely implemented for forecasting or predictions on statistics numbers i.e., when the result is an actual or nonstop value. it comes within Supervised Learning (SL), it is implemented with learned data to forecast latest trial data.

- **Classification:** A classification process, a technique of SL, presents a summary from detected data as one or many output in a class form.
- **Clustering:** Clustering is a ML approach that includes categorize data value into predefined class. If there are some items or objects, then clustering based approach is used to examine and classify them according to their features and similarity. It is unsupervised based approach which has not any labelled data and used for classification based on training datasets and form clusters on the behalf of matching of different features.

2.3Image Captioning and Deep Learning

Deep learning (DL) stimulates the human brain and motivated by the biological system just like human. It uses input layer and combination of hidden layers which work as an object detection and dimension reduction and finally output layer classify the object based upon their features and class (figure 4).

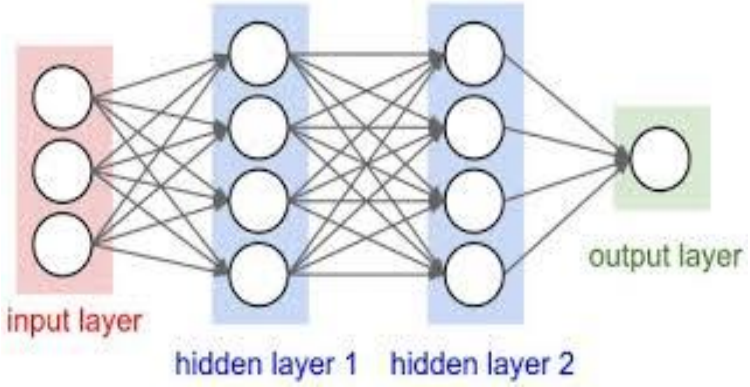

Figure 4: Structural Representation of a Deep Neural Network Comprising Multiple Hidden Layers for Hierarchical Feature Learning and Complex Pattern Recognition

The Encoder and Decoder architecture is utilized for a kind of setting where a variation of length of input-sequence of the sentence is mapped over the variation length provides out-sequence. The same model can also be trained for image description or classification. Encoder classify the image and decoder are used for description of the image. ConvNets is a type of CNN which has the multiple layers process the image and classify in their class. LSTMs and RNN are used for the generation of text from the word detected by the CNN. LSTM can store and retain the previous classification text for long time.

The Encoder and Decoder architecture (figure 5) is utilized for a kind of setting where a variation of length of input-sequence of the sentence is mapped over the variation length provides out-sequence. The same model can also be trained for image description or classification. Encoder classify the image and decoder are used for description of the image.

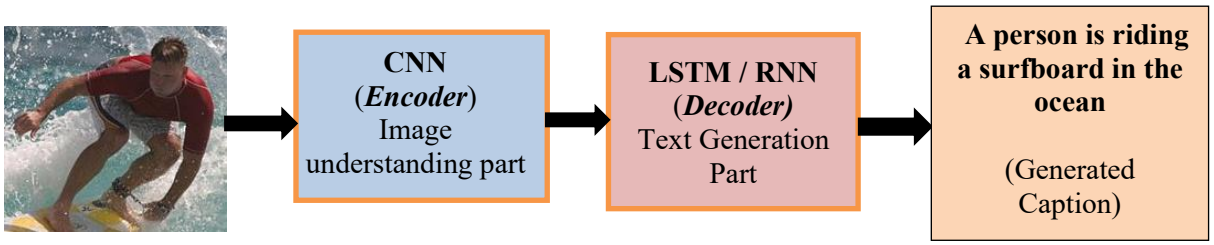

Figure 5: illustration of the Encoder-Decoder Architecture Used in Image Captioning

- **CNN (Convolutional Neural Network)**: A CNN is a specialized for applications in image and video recognition. It is mainly used in image analysis tasks like image recognition, Object detection and Segmentation [66] There are three types of layers in CNN Convolution Layer, Pooling Layer and Fully-Connected Layer.
- **VGG-16 Net:** VGG16 is a 16-layer architecture consists of convolution layers, pooling layer and at end fully connected layer. VGG network is a deeper networks with much smaller filters [12]. ResNet and DenseNet also most widely used for encoding the image [7][15][16].

Long Short-Term Memory Networks is a kind of Recurrent Neural Network. It is most widely used for the image captioning. LSTMs and RNN are used for the generation of text from the word detected by the CNN. LSTM can store and retain the previous classification text for long time [1][15][17][18][19][20][21][22][23][24][25]. RNN is less popular comparative to the LSTM. In RNN inputs and

results are not dependent to each other but they show dependency if there is a need to forecast the word of a sentence [15][26][27] [28][29].

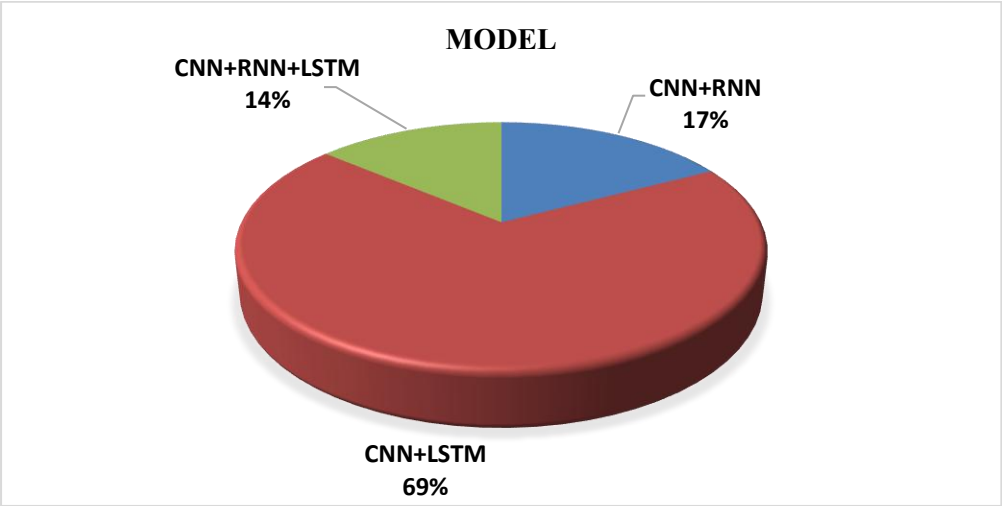

Figure 6: Summary of Deep Learning Models used in Literature for Image Captioning

In conclusion CNN represents the better results with LSTM (figure 6). RNN is a type of LSTM but it is not so popular as shown in figure 7.

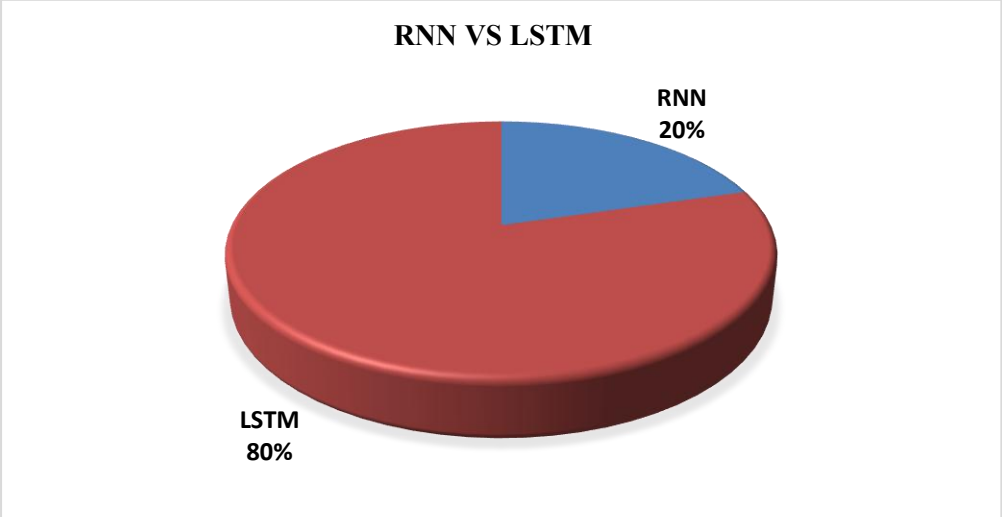

Figure 7: Summary of RNN Vs LSTM Models used in Literature for Image Captioning

2.4Transfer Learning in Image Captioning

Traditional learning is isolated and occurs purely based on specific tasks, datasets and training separate isolated models on them (figure 8). No knowledge is retained which can be transferred from one model to another. Thus, the key motivation, especially considering the context of deep learning is the fact that most models which solve complex problems need a whole lot of data, and getting vast amounts of labeled data for supervised models can be really difficult, considering the time and effort it takes to label data points. A simple example would be the ImageNet dataset, which has millions of images pertaining to different categories [30].

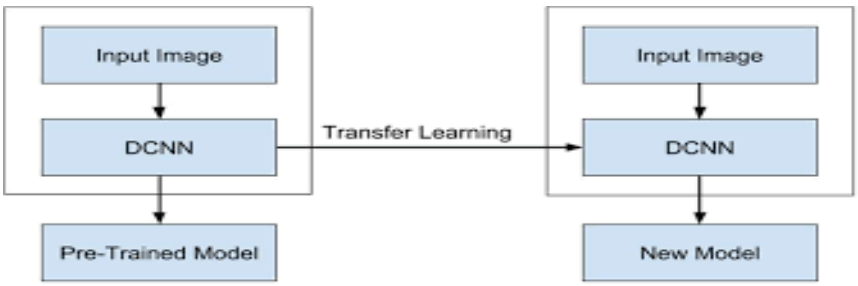

Figure 8: Illustration of Transfer Learning in Image Captioning

However, getting such a dataset for every domain is tough. Besides, most deep learning models are very specialized to a particular domain or even a specific task. While these might be state-of-the-art models, with really high accuracy and beating all benchmarks, it would be only on very specific datasets and end up suffering a significant loss in performance when used in a new task which might still be similar to the one it was trained on. This forms the motivation for transfer learning, which goes beyond specific tasks and domains, and tries to see how to leverage knowledge from pre-trained models and use it to solve new problems. In transfer learning, you can leverage knowledge (features, weights etc) from previously trained models for training newer models and even tackle problems like having less data for the newer task [30].

2.4 Use of Data Augmentation in Image Captioning

Data Augmentation is a method which can be used to enhance the size of the dataset by using various effects like rotation, scaling etc (figure 9). There is the problem of small dataset in most of the domain specific real-world problems like medical data, students' data etc. Data Augmentation is very helpful to increase the size dataset and train the mode. In over-all, having a large dataset is important for both ML and DL models [31]. We can augment Audio, Text, Images, Any other types of data.

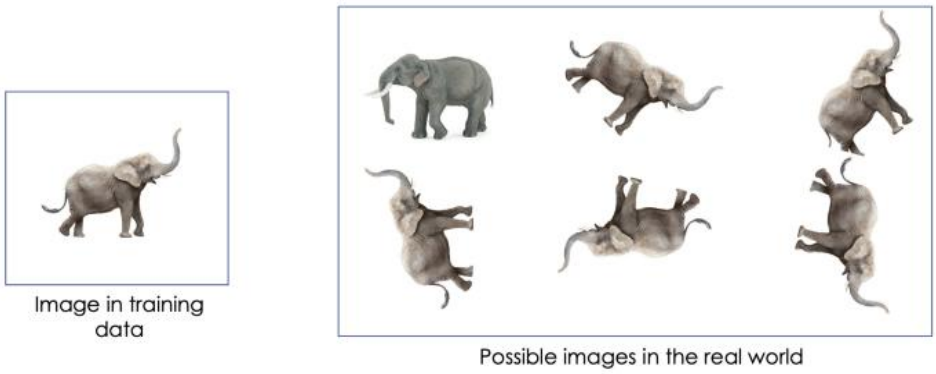

Figure 9: Impact of Data Augmentation Techniques on Image Captioning

3. STATE OF THE ART DEEP LEARNING BASED IMAGE CAPTIONING

The global features of the image are mined by the deep CNN model and semantic features learned by LSTM. Three datasets grey image, Twitter and Flicker are tested with the DMAF model and improved the semantic analysis accuracy of Flicker image dataset. The data-driven approaches are also used for the image description generation. In [19][40], retrieval-based techniques are used to develop the data-driven approaches of image description generation. The two methods are developed and evaluated to map photographs to natural language descriptions automatically.

A Hybrid model based on CNN and LSTM is proposed for the Indonesian image caption dataset named as FEEH-ID[38].The size of the image training is a huge problem of the nearest neighbour based image annotation. The selection of the image samples is a limitation of NN based image annotation. Effective prototyping of training images is achieved by the genetic algorithm [52]. The NN based algorithms present the optimal outcomes for automatic image annotation.

A bidirectional framework for automatic image description using the densely connected convolutional neural network is developed in [14].The Dense CNN model is used for encoding and BiLSTM performed decoding task. The forward and backward analysis of trained captions is provided in context of image description. A Game theoretic search approach is also implemented for word and best caption selection. The BLEU score is higher in case of proposed Dense CNN with BiLSTM model for image description generation. Similarly, a modular dense captioning architecture based on CNN and LSTM model is proposed for detailed image caption generation. The results are evaluated in terms of BLEU, ROUGE-L and METEOR. The sentences are generated by the combination of region attributes and object attributes. In [16] another dataset MSCOCO is also tested by the CNN-RNN features based image caption description. A regional object detector, RNN based attribute prediction and encoder decoder language generator embedded with two RNN is proposed to produce detailed description of a given image[5]. The proposed model is tested on the IAPR-TC 12 dataset. The proposed embedded model provided superior performance in case of cross-domain indoor dataset. A new approach of image annotation by combination of label generation, textual attention mechanism and image description generation with CNN-LSTM encoder decoder is proposed for flickr8K dataset [2].

Most of the researchers have focused on the image/text features extraction. For the second part of the work, they followed the conventional machine learning/deep learning algorithms. The semantic analysis between the image and their descriptive text from the dataset has been little attended. The idea of semantic relationship features extraction of stuff in the image with the other objects in the same image fills the research gap[13].The frequency of occurrence of text in the image description is also correlated by the action in the image by very few researchers.

Table 1: The summary of Datasets, Model and Performance of different approaches

Authors/year	Datasets	Models	Performance
Khurram et al. (2021) [17], " Dense-CaptionNet: a Sentence Generation Architecture for Fine-grained Description of Image Semantics"	MSCOCO and IAPR TC-12 used for detailed description	CNN-RNN based	B1: 0.128, B2: 0.064, B3: 0.031, B4: 0.016, R-L: 0.216, M: 0.070 **Remarks:** Complex scene detailed description, **Limitation:** • Overfitting • Not capturing inter-object relationship • Degraded accuracy when the image is rotated. • Domain specific description not generated.
Yao et al. [2021] [2] "Deep neural network compression through interpretability-based filter pruning"	CIFAR-10, ImageNet	Compressed DNN based on VGG-16, ResNet50,	**Remarks:** • Under CIFAR With 60% compression rate of the VGG-16 accuracy=.8429 (accuracy dropped only .0322) and storage space can be compressed to 9.42 Mb. • On ResNet50 with ImageNet better than other pruning methods • Used single layer filter pruning method
Zhang et al. [2020] [14], " Image Captioning via semantic element embedding"	MSCOCO, Flickr 8K, Flickr 30K	CNN-LSTM, ELSTM	B1:75.7, B2: 59.4, B3: 45.3, B4: 34.6, M: 26.8, C: 109.6, R: 56.0 **Remarks:** • Integrate the local and global objects. • Generate semantic features based on local and global description.
Liu et al. (2020) [3], " Image caption generation with duel attention mechanism"	AIC-ICC (Chinese's caption benchmark dataset)	CNN-LSTM	B1:0.572, B2:0356, B3: 0.293, B4: 0.229, M: 0.297, R: 0.431, C: 0.613 **Remarks:** • Duel (visual and textual) attention image caption generation model **Advantage:** More integrity of text image and labels fully generated
Zhang et al. (2019) [46]	Sydney and RSICD	CNN-LSTM based	B1: 0.8143, B2: 0.7351, B3: 0.6586, B4: 0.5806, M: 0.4111, R: 0.7195, C: 2.3021. **Advantages:** Provide robust performance in semantic description of image. **Limitation:** Tested on high resolution of image
Jin et al. (2019) [51]	Core15k, ESP Game IAPR TC-12	VAM-CRF (based on SVM)	Core15k: F1: 0.497, N^+: 199 ESP Game: F1: 0.418, N^+: 259

			IAPR TC-12: F1: 0.451, N^+: 281 **Advantages:** Non-salient region drawback improved
Huang et al. (2019) [47]	Getty Image, Twitter Flickr-w, Flickr-m	DMAF based LSTM	Getty: P: 0.882, Recall: 0.851, F1:0.866, Accuracy: 0.869 Twitter: 0.778, 0.760, 0.769, 0.763 Flickr-w: 0.855, 0.845, 0.850, 0.859 Flickr-m: 0.882, 0.870, 0.876, 0.880 **Advantage:** Visual and Semantic information prediction **Drawback:** Complex
Xiao et al. (2019) [49]	*MSCOCO, Flickr30k	End to end Deep CNN With ML-WGAN model	B1:0.753, B2: 0.596, B3: 0.457, B4: 0.347, M: 0.262, R: 0.556, C: 0.106 **Advantage:** Multi-label data augmented **Drawback:** Can't be used for small scale database.
Li, J., Yao (2019) [52]	MSCOCO	CNN-LSTM	B1:0.81, B2: 0.659, B3: 0.515, B4: 0.395, M: 0.293, C: 0.1309, R: 0.589 Latter achieved double attention under the influence of the visual information
Herdade S. (2019) [50]	MSCOCO	Object Relation Transformer (Geometric features +Global Features)	B1:80.5, B4: 38.6, C: 128.3, S: 22.6, M: 28.7, R: 58.4 improved interpretability of the model
Mulyanto et al. (2019) [40]	FEEH-ID for Flickr	CNN-LSTM	B1: 0.473, B2:0.339, B3: 0.231, B4 :0.153 **Advantage**: Multiple language image description achieved
Kinghorn et al. (2019) [27]	Flickr 8K, Flickr 30K MSCOCO	Supervised and unsupervised deep learning methods survey	B1:0.094, B2: 0.046, B3: 0.034, B4: 0.013, M: 0.059, R: 0.205, SPICE: 0.06 **Advantage**: Survey
SR et al. (2019) [15]	Flickr 8K	Dense CNN model with BiLSTM	B1: 0.699, B2:0.563, B3:0.4645, B4:0.4295 grammatical correctness of the description
Khang et al. (2019) [36]	Flickr8K	LSTM	B1 :0.50827, B2: 0.272958 B3: 0.155084, B4: 0.084381, M: 0.21444, C:0.24516, R: 0.43794 **Advantage**: Description generation errorless
Baig, M.M.A. et al. (2019) [33]	Flickr30K	VGGNet, GoogleNet, ResNet, LSTM	B1: 63, B2: 42.1, B3 :24.2, B4:14.6, M: 42.3 C: 25, improve accuracy
Yu, N. et al. (2019) [13]	*MSCOCO, **Flickr30K	GoogleNet, LSTM	*B1: 74.0, B2: 56.7, B3: 43.3, B4: 31.3, M:25.5, C: 98.3 ** B1: 64.6, B2: 43.8, B3 :31.9, B4:22.4, M: 19.2, C: 39.6
He, X. et al. (2019) [23]	Flickr30K	VGGNet, LSTM	B1:63.8, B2: 44.6, B3: 30.7, B4: 21.1
Wu, Q. et al. (2018) [18]	*MSCOCO, **Flickr30K	VGGNet, LSTM	*B1: 74, B2: 56, B3: 42, B4: 31, M:26 ** B1: 73, B2: 55, B3 :40, B4:28,

Anderson, P et al. (2018) [37]	*MSCOCO, IAPR TC	R-CNN	B1: 0.767, B2: 0.601, B3: 0.449, B4: 0.294, M: 0.254, R: 0.539
Chang, Y.S. (2018) [18]	*MSCOCO, **Flickr30K	VGGNet, LSTM	*B1: 72.5, B2: 51, B3: 36.2, B4: 25.9, M:24.5 ** B1: 68.4, B2: 45.5, B3 :31.3, B4:21.4 M: 19.9
Cornia, M. et al. (2018) [55]	*MSCOCO, **Flickr30K	ResNet, LSTM	*B1: 70.8, B2: 53.6, B3: 39.1, B4: 28.4, M:24.8, C: 89.8 ** B1: 61.5, B2: 43.8, B3 :30.5, B4:21.3, M: 20, C: 46.4
Zhu, X. et al. (2018) [10]	MSCOCO	VGGNet, ResNet, LSTM	B1: 74.2, B2: 57.7, B3: 43.8, B4: 33, C: 105.8
Ge et al. (2018) [53]	MSCOCO	CNN	Pre:0.38, Rec:0.39, F1:0.392, F1:0.168
Maihami et al. (2018) [54]	Core15k ,IAPR TC12 and MIR Flickr datasets	GENMF based KNN approach	Pre:0.23, Recall:0.27, F-score:0.25 **Advantage**: Improved Accuracy
Kinghorn et al. (2018) [6]	ImageNet, PubFig MSCOCO	Hierarchical deep Learning scheme(R-CNN and RNN)	B1:0.231, B2: 0.099, B3: 0.046, B4: 0.024, M: 0.067, R: 0.183
Wang, C., (2018) [11]	Flickr8K Flickr30K MSCOCO	CNN-LSTM	B: 0.67 M: 19.5 C: 66.0 **Remarks:** highly competitive performance without integrating additional mechanism Data augmentation techniques are used.
Shi and Zou (2017) [56]	Google Earth, GaoFen-2	Fully convolutional networks	**Precision:** 95.3% Recall: 94.1% high-resolution optical images
Tariq and Foroosh (2017) [57]	TIME Magazine	Context Driven Framework	METEOR: 0.053 TER: 1.75 **Advantage**: importance of weighted auxiliary information **Disadvantage**: only annotations no captions
Yuan, A et al. (2017)][41]	*MSCOCO, **Flickr30K	VGGNet, LSTM	*B1: 70.1, B2: 50.2, B3: 35.8, B4: 25.5, M:24.1 ** B1: 67.9, B2: 44, B3 :29.2, B4:20.9, M: 19.7
Lintas, A. et al. (2017) [28]	MSCOCO	CNN-LSTM	B4: 31.1, C: 93.2
Venugopalan et al. (2017) [42]	MSCOCO	CNN-RNN	BLEU: 0.2132 ,F1:48.79
Fu, K. et al. (2017) [20]	*MSCOCO, **Flickr30K	ResNet, LSTM	*B1: 72.4, B2: 55.5, B3: 41.8, B4: 31.3, C: 95.5, M:24.8 ** B1: 64.9, B2: 46.2, B3 :32.4, B4: 22.4, C: 47.2, M: 19.4
Vinyals, O et al. (2017) [34]	*MSCOCO, **Flickr30K	GoogleNet, LSTM	*B4: 32.1, C: 99.8 ** B1: 66,
Dai, B. et al. (2017) [58]	MSCOCO	VGGNet, ResNet, DenseNet, LSTM, RNN	B1: 91, B2: 83.1, B3: 72.8, B4: 61.7, C: 102.9, M:35

Kun Fu et al. (2017) [20]	MSCOCO	CNN-LSTM	B1: 72.4, B2: 55.5, B3: 41.8, B4: 31.3, C: 95.5, M:24.8
Ordonez et al. (2016) [43]	Object Detection, Image Parsing, Caption Parsing Scene classification	Global image descriptor	B: 0.1260, R: 0.2470 **Advantage:** Large scale category image description
Karpathy A et al. (2015) [35]	*MSCOCO, IAPRTC-12	CNN-LSTM	B1:0.730, B2: 0.530, B3: 0.393, B4: 0.347, M: 0.241, R: 0.520
B1: BLEU1, B2: BLEU2, B3: BLEU3, B4: BLEU4, M: METEOR, R: ROUGE, C: CIDEr			

4. IMAGE CAPTIONING DATASETS

There are lots of data sets available for image captioning which are presents in the literature but MS COCO and Flicker data sets are very popular data sets. MS COCO is best suited data set as it also contains the set of non-iconic images. Our literature survey shows that 55% authors prefer MS COCO, 42% Flicker and 3% uses other data sets (figure 10).

- **MS COCO:** it one of the largest datasets with 330K images used for image captioning. Here in this dataset, there are five captions are associated with each image. It has eighty (80) objects and ninety-one (91) stuffs classes. It is blessed with millions of objects class and more than two lac people with main features [10] [11][33] [34][35].
- **Flickr8k/30k**: It is a dataset extended from Flickr 8k. It consists of 30,000 images that are paired with five different captions. The images in the dataset contains human involved in everyday activities and events [36][37].

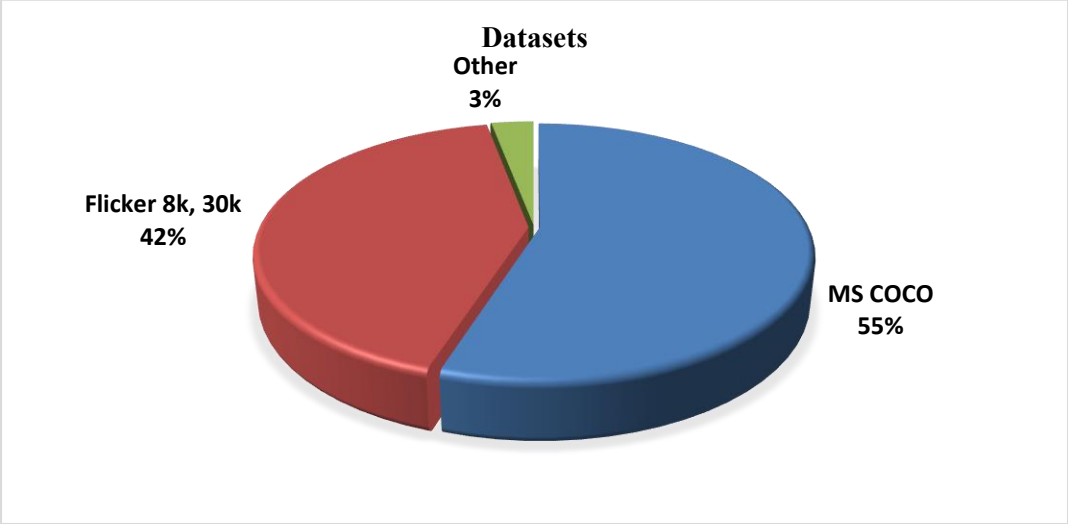

Figure 10: Percentage of datasets used for image captioning in studies.

5. IMAGE CAPTIONING EVALUATION METRICS

In image captioning final result or captions are evaluated by using different evaluation metrics like BLEU, ROUGH-L, CIDEr, METEOR, and SPICE. As shown in figure 11. it is shows that BLEU is most popular and SPICE is least one in the literature. Further in the initial stage precision, recall and F-score also used to evaluate the same.

a) **Precision:** the proportion of positive cases that were correctly identified.
b) **Recall:** the proportion of actual positive cases that are correctly identified
c) **F-score:** F1 Score is the harmonic mean for precision and recall values. The formula for F1 score goes this way-

$$F1 = 2 * (Precision * Recall) / (Precision + Recall) \quad (1)$$

d) **BLEU:** BLUE stands for Bilingual Evaluation Understudy which is used to measure quality of generated caption. It is a metric for evaluating a generated sentence to a reference sentence. The perfect match is 1.0 and a perfect mismatch is 0.0. It works only for short captions but not long paragraphs.[15][27][33][38][39] [40] .

e) **METEOR:** METEOR stands for metric for evaluation and translation with explicit ordering. BLEU consider of whole text produced intense the rank of every and separated sentence produced the METEOR. It improves precision and recall value[38].

$$\mathbf{Fmean} = \frac{\mathbf{10PR}}{\mathbf{R+9P}} \quad (2)$$

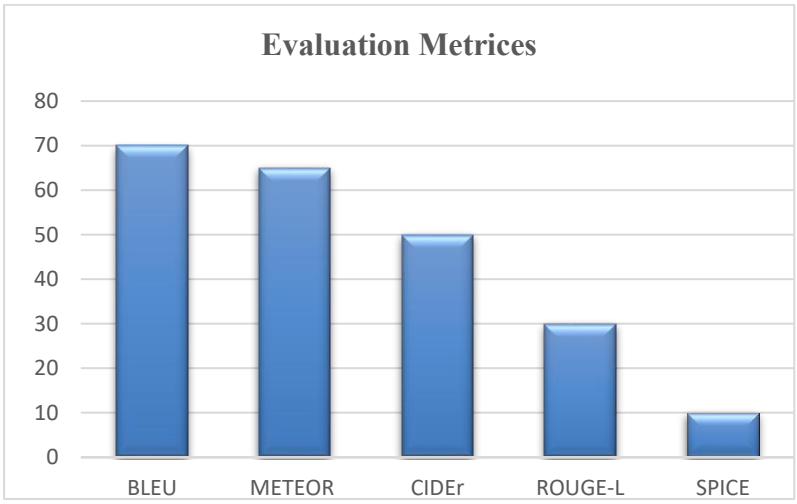

Figure 11: Evaluation Metrics used in literature

6. Challenges In Image Captioning

Image captioning faces the following challenges:

- To generate complete natural language sentences like a human being;
- To make the generated sentence grammatically correct;
- To make the caption semantics as clear as possible and consistent with the given image content.
- To enhance the adaptability of the network to work according to different situations and concerns.
- Sentiment-based annotations is a major challenge as not many datasets are available for analyse the sentiments.

7. Latest Trends In Image Captioning

The latest trends in automatic image captioning reflect a shift toward more intelligent, flexible, and human-like systems, powered primarily by advances in deep learning and multimodal AI. Vision-language transformers such as BLIP, OFA, and GIT have outperformed traditional CNN-RNN models by leveraging cross-modal attention and large-scale pretraining. Additionally, multimodal foundation models like GPT-4V and Gemini now perform captioning, visual question answering, and reasoning in a single architecture without task-specific tuning. Prompt-based and zero-shot models such as BLIP-2 and CLIP allow captioning of unseen images with natural language instructions, enhancing adaptability. There is also a growing focus on emotion-aware and personalized captioning that incorporates user context, along with multilingual generation for global accessibility. In parallel, video captioning is gaining momentum by modeling temporal relationships across frames, while explainable and controllable captioning ensures transparency and user-guided outputs. To support deployment in edge devices, lightweight models using pruning and quantization are being developed for real-time performance. Furthermore, self-supervised learning and synthetic data are being used to reduce annotation costs and improve generalization, making modern image captioning more scalable and robust than ever before.

7.1Vision-Language Transformers

Recent advancements in image captioning have been largely driven by transformer-based architectures that jointly process visual and textual modalities. Unlike traditional CNN-RNN frameworks, models like BLIP, OFA, and GIT employ cross-modal attention to extract global semantic relationships, resulting in more coherent and contextually accurate captions. These models have become state-of-the-art due to their scalability and ability to pretrain on large datasets, outperforming older architectures in most benchmark datasets.

7.2Multimodal Foundation Models

Multimodal foundation models such as GPT-4V, Gemini, and Claude Vision integrate vision and language understanding into a single, powerful framework. These models are capable of performing various tasks—including captioning, VQA, and image reasoning—without needing fine-tuning. Their versatility allows them

to generate human-like captions for unseen images by leveraging a unified knowledge base trained across multiple modalities.

7.3Prompt-Based and Zero-Shot Captioning

A growing trend is the use of prompt-based models that can perform captioning without task-specific training. Models like BLIP-2 and CLIP can interpret images and respond to textual prompts in a zero-shot manner, making them highly adaptable. This flexibility allows these models to generate captions for unseen categories or datasets, significantly reducing the need for large-scale labeled training data.

7.4Personalized and Context-Aware Captioning

Newer image captioning models are being designed to include personalization and context-awareness. For instance, captions can be adapted based on user preferences, location, or the application domain. This is especially important in assistive technologies, where personalized captioning enhances usability for visually impaired users or tailored social media content generation.

7.5Emotion and Sentiment-Aware Captioning

Traditional captioning models focus on object and action recognition, but emerging models are now integrating emotional intelligence. By recognizing facial expressions, body language, or contextual clues, models can generate emotionally resonant captions such as "a joyful child playing in the park" which enhance the richness and relatability of descriptions.

7.6Multilingual Caption Generation

The ability to generate captions in multiple languages is becoming essential in global applications. Modern architectures incorporate multilingual transformers like mBART and XLM-R to generate captions in English, Hindi, French, and more. This trend facilitates accessibility and usability across diverse linguistic audiences without needing separate models for each language.

7.7Video and Temporal Captioning

While static image captioning remains important, the shift toward video captioning has introduced new challenges and opportunities. Temporal attention mechanisms now allow models to track object movements and changes across frames, enabling dynamic storytelling in real-time. This is critical for applications in surveillance, autonomous driving, and media content generation.

7.8Explainable and Controllable Captioning

With growing concerns about AI transparency, there is a demand for captioning models that offer interpretability. Explainable AI (XAI) techniques are being applied to visualize which image regions influenced specific words in a caption. Additionally, controllable captioning allows users to guide the model's focus—whether stylistically (e.g., formal, humorous) or thematically (e.g., focus on people vs. objects).

7.9Real-Time Edge Deployment

To support deployment in mobile devices, AR/VR systems, and IoT devices, researchers are optimizing captioning models using model compression techniques like pruning, quantization, and knowledge distillation. This trend is enabling real-time image captioning with minimal latency, crucial for applications like smart glasses and wearable AI assistants.

7.10 Synthetic Data and Self-Supervised Learning

To overcome the limitations of manually annotated datasets, synthetic data generation and self-supervised learning have gained popularity. Techniques such as masked image modeling and contrastive learning allow models to learn effective representations without explicit supervision. This not only reduces the reliance on expensive human-labeled data but also boosts model performance on rare or unseen scenarios.

8. CONCLUSION

Latest advancements in technology allow computer systems to generate captions of social media images. Often, these descriptions are very useful for visually impaired people which make them able to use the digital platform. Image captioning also helpful in describing the scene by converting the view into text which may further useful for automatic vehicle and disabled people. It also might help Google Image search. There are many models that have already been presented to generate meaningful captions for images. These models are quite good, but have some constraints. Image captioning still have a long way to go in improving the accuracy of captioning the events in images. We reviewed some of the recent deep learning-based works, and it is hard to compare different works due to the different combination of structures, using different parameters and implying various datasets. We also noticed that there is a lot of room for improvement in accuracy.

REFERENCES:

[1] M. D. Zakir Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, "A comprehensive survey of deep learning for image captioning," ACM Comput. Surv., vol. 51, no. 6, 2019, doi: 10.1145/3295748.
[2] D. Yu, Q. Xu, H. Guo, C. Zhao, Y. Lin, and D. Li, "An efficient and lightweight convolutional neural network for remote sensing image scene classification," Sensors (Switzerland), vol. 20, no. 7, 2020, doi: 10.3390/s20071999.
[3] M. Liu, L. Li, H. Hu, W. Guan, and J. Tian, "Image caption generation with dual attention mechanism," Inf. Process. Manag., vol. 57, no. 2, p. 102178, 2020, doi: 10.1016/j.ipm.2019.102178.
[4] Komal Vij and Yaduvir Singh, "Enhancement of Images Using Histogram Processing Techniques," Int. J. Comp. Tech. Appl., vol. Vol 2, no. 2, pp. 309–313, 2011, [Online]. Available: http://www.ijcta.com/documents/volumes/vol2issue2/ijcta2011020212.pdf.
[5] J. Shen et al., "Unified structured learning for simultaneous human pose estimation and garment attribute classification," IEEE Trans. Image Process., vol. 23, no. 11, pp. 4786–4798, 2014, doi: 10.1109/TIP.2014.2358082.
[6] P. Kinghorn, L. Zhang, and L. Shao, "A region-based image caption generator with refined descriptions," Neurocomputing, vol. 272, pp. 416–424, 2018, doi: 10.1016/j.neucom.2017.07.014.
[7] W. Wang, Y. Li, T. Zou, X. Wang, J. You, and Y. Luo, "A novel image classification approach via dense-mobilenet models," Mob. Inf. Syst., vol. 2020, 2020, doi: 10.1155/2020/7602384.
[8] G. T. U. A. Colleges et al., "Microsoft COCO," Eccv, no. June, pp. 740–755, 2014.
[9] P. F. Felzenszwalb, R. B. Girshick, D. Mcallester, and D. Ramanan, "Object Detection With Partbase," IEEE Trans. Pattern Anal. Mach. Intell., vol. 32, no. 9, pp. 1627–1645, 2010.
[10]X. Zhu, L. Li, J. Liu, Z. Li, H. Peng, and X. Niu, "Framing image description as a ranking task: Data, models and evaluation metrics," Neurocomputing, vol. 319, pp. 55–65, 2018, doi: 10.1016/j.neucom.2018.08.069.
[11]C. Wang, H. Yang, and C. Meinel, "Image Captioning with Deep Bidirectional LSTMs and Multi-Task Learning," ACM Trans. Multimed. Comput. Commun. Appl., vol. 14, no. 2s, pp. 988–997, 2018, doi: 10.1145/3115432.
[12]Z. Gan et al., "Semantic Compositional Networks for Visual Captioning : Supplementary Material," Proc. - 2017 IEEE Winter Conf. Appl. Comput. Vision, WACV 2017, pp. 1–3, 2017.
[13]N. Yu, X. Hu, B. Song, J. Yang, and J. Zhang, "Topic-Oriented Image Captioning Based on Order-Embedding," IEEE Trans. Image Process., vol. 28, no. 6, pp. 2743–2754, 2019, doi: 10.1109/TIP.2018.2889922.
[14]X. Zhang, S. He, X. Song, R. W. H. Lau, J. Jiao, and Q. Ye, "Image captioning via semantic element embedding," Neurocomputing, vol. 395, no. xxxx, pp. 212–221, 2020, doi: 10.1016/j.neucom.2018.02.112.
[15]S. R. Sreela and S. M. Idicula, "Dense model for automatic image description generation with game theoretic optimization," Inf., vol. 10, no. 11, 2019, doi: 10.3390/info10110354.
[16]X. He, Y. Yang, B. Shi, and X. Bai, "VD-SAN: Visual-Densely Semantic Attention Network for Image Caption Generation," Neurocomputing, vol. 328, pp. 48–55, 2019, doi: 10.1016/j.neucom.2018.02.106.
[17]I. Khurram, M. M. Fraz, M. Shahzad, and N. M. Rajpoot, "Dense-CaptionNet: a Sentence Generation Architecture for Fine-grained Description of Image Semantics," Cognit. Comput., vol. 13, no. 3, pp. 595–611, 2021, doi: 10.1007/s12559-019-09697-1.
[18]Q. Wu, C. Shen, P. Wang, A. Dick, and A. Van Den Hengel, "Image Captioning and Visual Question Answering Based on Attributes and External Knowledge," IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 6, pp. 1367–1381, 2018, doi: 10.1109/TPAMI.2017.2708709.
[19]R. Staniute and D. Šešok, "A systematic literature review on image captioning," Appl. Sci., vol. 9, no. 10, 2019, doi: 10.3390/app9102024.
[20]K. Fu, J. Jin, R. Cui, F. Sha, and C. Zhang, "Aligning Where to See and What to Tell: Image Captioning with Region-Based Attention and Scene-Specific Contexts," IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 12, pp. 2321–2334, 2017, doi: 10.1109/TPAMI.2016.2642953.
[21]A. Shin, Y. Ushiku, and T. Harada, "Image captioning with sentiment terms via weakly-supervised sentiment dataset," Br. Mach. Vis. Conf. 2016, BMVC 2016, vol. 2016-Septe, pp. 53.1-53.12, 2016, doi: 10.5244/C.30.53.
[22]K. S. Tai, R. Socher, and C. D. Manning, "Improved semantic representations from tree-structured long short-Term memory networks," ACL-IJCNLP 2015 - 53rd Annu. Meet. Assoc. Comput. Linguist. 7th Int. Jt. Conf. Nat. Lang. Process. Asian Fed. Nat. Lang. Process. Proc. Conf., vol. 1, pp. 1556–1566, 2015, doi: 10.3115/v1/p15-1150.
[23]X. He, B. Shi, X. Bai, G. S. Xia, Z. Zhang, and W. Dong, "Image Caption Generation with Part of Speech Guidance," Pattern Recognit. Lett., vol. 119, no. February 2018, pp. 229–237, 2019, doi: 10.1016/j.patrec.2017.10.018.
[24]C. Gan, Z. Gan, X. He, J. Gao, and L. Deng, "StyleNet: Generating attractive visual captions with styles," Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017, vol. 2017-Janua, pp. 955–964, 2017, doi: 10.1109/CVPR.2017.108.
[25]T. Chen et al., "'Factual' or 'emotional': Stylized image captioning with adaptive learning and attention," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 11214 LNCS, pp. 527–543, 2018, doi: 10.1007/978-3-030-01249-6_32.
[26]J. Donahue et al., "Long-term recurrent convolutional networks for visual recognition and description," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 07-12-June, pp. 2625–2634, 2015, doi: 10.1109/CVPR.2015.7298878.
[27]P. Kinghorn, L. Zhang, and L. Shao, "A hierarchical and regional deep learning architecture for image description generation," Pattern Recognit. Lett., vol. 119, pp. 77–85, 2019, doi: 10.1016/j.patrec.2017.09.013.
[28]S. Otte, T. Schmitt, K. Friston, and M. V. Butz, "Inferring adaptive goal-directed behavior within recurrent neural networks," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 10613 LNCS, pp. 227–235, 2017, doi: 10.1007/978-3-319-68600-4_27.
[29]M. Tanti, A. Gatt, and K. P. Camilleri, "Where to put the image in an image caption generator," Nat. Lang. Eng., vol. 24, no. 3, pp. 467–489, 2018, doi: 10.1017/S1351324918000098.
[30]M. T. Rosenstein, Z. Marx, L. P. Kaelbling, and T. G. Dietterich, "To TUI or not to TUI," NIPS 2005 Work. Transf. Learn., vol. 898, p. 3, 2005.
[31]C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," J. Big Data, vol. 6, no. 1, 2019, doi: 10.1186/s40537-019-0197-0.

[32]A. Nasif, Z. A. Othman, and N. S. Sani, "The deep learning solutions on lossless compression methods for alleviating data load on iot nodes in smart cities," Sensors, vol. 21, no. 12, 2021, doi: 10.3390/s21124223.
[33]M. M. A. Baig, M. I. Shah, M. A. Wajahat, N. Zafar, and O. Arif, "Image Caption Generator with Novel Object Injection," 2018 Int. Conf. Digit. Image Comput. Tech. Appl. DICTA 2018, pp. 1–8, 2019, doi: 10.1109/DICTA.2018.8615810.
[34]O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and Tell: Lessons Learned from the 2015 MSCOCO Image Captioning Challenge," IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 4, pp. 652–663, 2017, doi: 10.1109/TPAMI.2016.2587640.
[35]A. Karpathy and L. Fei-Fei, "Deep Visual-Semantic Alignments for Generating Image Descriptions - Karpathy_Deep_Visual-Semantic_Alignments_2015_CVPR_paper.pdf," Cvpr, 2015, [Online]. Available: https://www.cv-foundation.org/openaccess/content_cvpr_2015/papers/Karpathy_Deep_Visual-Semantic_Alignments_2015_CVPR_paper.pdf%0Ahttp://www.cv-foundation.org/openaccess/content_cvpr_2015/papers/Karpathy_Deep_Visual-Semantic_Alignments_2015_CVPR_paper.pdf.
[36]Phyu Phyu Khaing | Mie Mie Aung | Myint San, "Natural Language Description Generation for Image using Deep Learning Architecture," Int. J. Trend Sci. Res. Dev., vol. 3, no. 5, pp. 1575–1581, 2019, doi: https://doi.org/10.31142/ijtsrd26708.
[37]Y. S. Chang, "Fine-grained attention for image caption generation," Multimed. Tools Appl., vol. 77, no. 3, pp. 2959–2971, 2018, doi: 10.1007/s11042-017-4593-1.
[38]M. Denkowski and A. Lavie, "Meteor Universal: Language Specific Translation Evaluation for Any Target Language," pp. 376–380, 2015, doi: 10.3115/v1/w14-3348.
[39]P. Anderson et al., "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 6077–6086, 2018, doi: 10.1109/CVPR.2018.00636.
[40]E. Mulyanto, E. I. Setiawan, E. M. Yuniarno, and M. H. Purnomo, "Automatic Indonesian Image Caption Generation using CNN-LSTM Model and FEEH-ID Dataset," 2019 IEEE Int. Conf. Comput. Intell. Virtual Environ. Meas. Syst. Appl. CIVEMSA 2019 - Proc., 2019, doi: 10.1109/CIVEMSA45640.2019.9071632.
[41]S. Venugopalan et al., "Captioning Images with Diverse Objects Supplementary Material," Cvpr, vol. 3, p. 8, 2017.
[42]M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," IJCAI Int. Jt. Conf. Artif. Intell., vol. 2015-Janua, pp. 4188–4192, 2015.
[43]V. Ordonez et al., "Large Scale Retrieval and Generation of Image Descriptions," Int. J. Comput. Vis., vol. 119, no. 1, pp. 46–59, 2016, doi: 10.1007/s11263-015-0840-y.
[44] and J. L. Quanzeng You1, Hailin Jin2, Zhaowen Wang2, Chen Fang2, "Image Captioning with Semantic Attention," J. Comput. Vis. Found., pp. 4651-4659,., 2016, doi: doi: 10.1109/CVPR.2016.503.
[45]K. Xu?, J. L. B. R. Kiros†, K. C. A. Courville?, R. S. R. S. Zemel†∗, and Y. Bengio?∗, "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention," Proc. 32nd Int. Conf. Mach. Learn., vol. 37, pp. 2048–2057, 2015.
[46]X. Zhang, X. Wang, X. Tang, H. Zhou, and C. Li, "Description Generation for Remote Sensing Images Using Attribute Attention Mechanism," Remote Sens., vol. 11, no. 6, p. 612, 2019, doi: 10.3390/rs11060612.
[47]F. Huang, X. Zhang, Z. Zhao, J. Xu, and Z. Li, "Image–text sentiment analysis via deep multimodal attentive fusion," Knowledge-Based Syst., vol. 167, pp. 26–37, 2019, doi: 10.1016/j.knosys.2019.01.019.
[48]G. Kulkarni et al., "Baby talk: Understanding and generating simple image descriptions," IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 12, pp. 2891–2903, 2013, doi: 10.1109/TPAMI.2012.162.
[49]X. Xiao, L. Wang, K. Ding, S. Xiang, and C. Pan, "Dense semantic embedding network for image captioning," Pattern Recognit., vol. 90, pp. 285–296, 2019, doi: 10.1016/j.patcog.2019.01.028.
[50]S. Herdade, A. Kappeler, K. Boakye, and J. Soares, "Image captioning: Transforming objects into words," Adv. Neural Inf. Process. Syst., vol. 32, no. NeurIPS, pp. 1–11, 2019.
[51]C. Jin, Q. M. Sun, and S. W. Jin, "A hybrid automatic image annotation approach," Multimed. Tools Appl., vol. 78, no. 9, pp. 11815–11834, 2019, doi: 10.1007/s11042-018-6742-6.
[52]J. Li, P. Yao, L. Guo, and W. Zhang, "Boosted transformer for image captioning," Appl. Sci., vol. 9, no. 16, pp. 1–15, 2019, doi: 10.3390/app9163260.
[53]H. Ge, Z. Yan, J. Dou, Z. Wang, and Z. Q. Wang, "A Semisupervised Framework for Automatic Image Annotation Based on Graph Embedding and Multiview Nonnegative Matrix Factorization," Math. Probl. Eng., vol. 2018, 2018, doi: 10.1155/2018/5987906.
[54]V. Maihami and F. Yaghmaee, "A genetic-based prototyping for automatic image annotation," Comput. Electr. Eng., vol. 70, pp. 400–412, 2018, doi: 10.1016/j.compeleceng.2017.03.019.
[55]M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "Paying More Attention to Saliency," ACM Trans. Multimed. Comput. Commun. Appl., vol. 14, no. 2, pp. 1–21, 2018, doi: 10.1145/3177745.
[56]Z. Shi and Z. Zou, "Can a Machine Generate Humanlike Language Descriptions for a Remote Sensing Image?," IEEE Trans. Geosci. Remote Sens., vol. 55, no. 6, pp. 3623–3634, 2017, doi: 10.1109/TGRS.2017.2677464.
[57]A. Tariq and H. Foroosh, "A Context-Driven Extractive Framework for Generating Realistic Image Descriptions," IEEE Trans. Image Process., vol. 26, no. 2, pp. 619–632, 2017, doi: 10.1109/TIP.2016.2628585.
[58]B. Dai and D. Lin, "Contrastive learning for image captioning," Adv. Neural Inf. Process. Syst., vol. 2017-Decem, no. Cl, pp. 899–908, 2017.